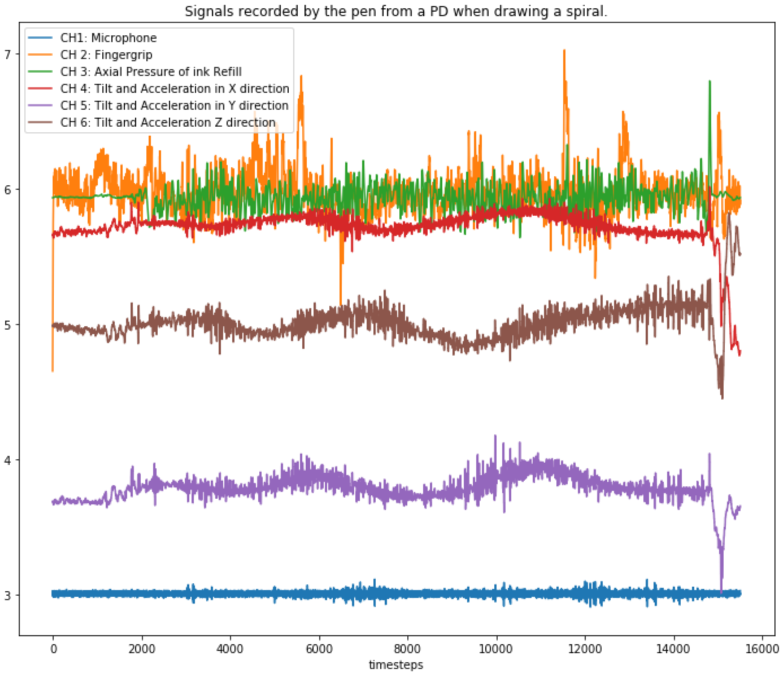
Report #6
PD-Internship
Lerner Paul
14/05/2018

This report is about my work on the NewHandPD dataset of Pereira et al. please refer to report #2 for analysis about related works on this dataset. Some of this report was originally present in the report on the code. I split it for better readability.

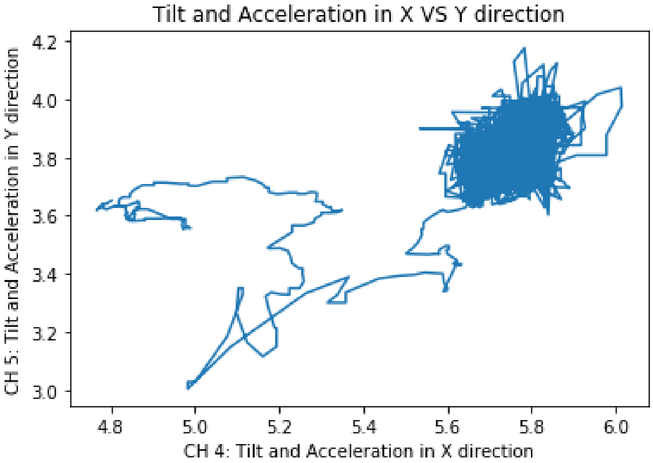
Summary

1 Data Exploration

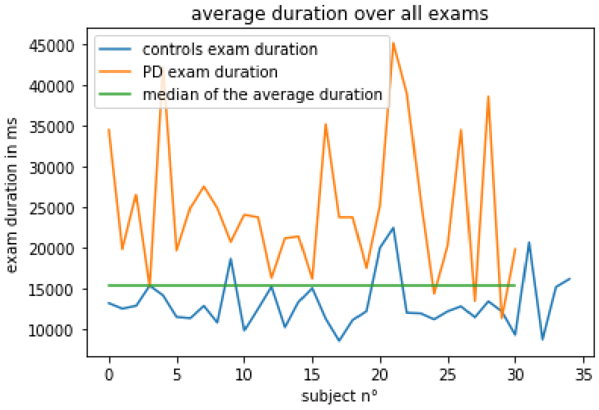
Hoping to achieve transfer learning I took a look at the NewHandPD database of Pereira et al. (cf. Report #2). However I'm not sure we'll be able to achieve transfer learning between NewHandPD and PaHaW as NewHandPD doesn't record the x and y coordinate but the "tilt and acceleration" in both x and y directions.

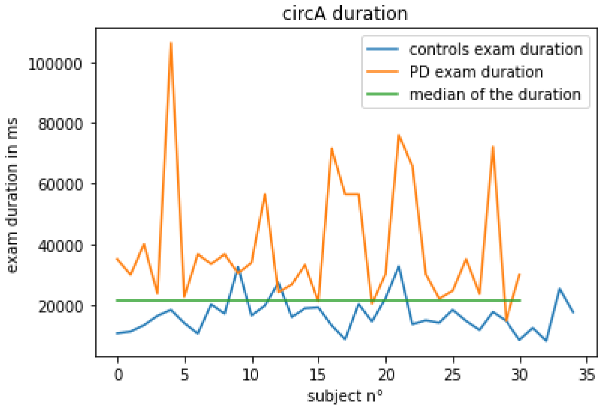


Thus the plot of X vs Y is not visually interpretable.



This dataset has a serious defect : the PDs exams take significantly much time than Controls' exams : 11.5 seconds in average ! Thus when plotting the average duration of the exams per subject we can clearly see a threshold that separates almost all controls and all PDs. The phenomenon is visible on every single task, see below for the CircA exam.





Thus when using a simple rule :

- subject is PD if exam_duration > 15242 ms *(median length over all subjects)*
- else control

I'm able to achieve 85% accuracy ! See 2 Classification for proper classification tests (i.e. with train-test etc.). See 3 Interpretation for interpretations of this defect.

2 Data analysis and classification

2.1 Statistical analysis

I used Spearman's rank correlation coefficient measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. I used SciPy's implementation to measure the correlation between the task duration and the targets. As you can see below, all the tasks' duration is positively correlated to the subject's target with a very strong significance (cf. p-value).

Table. Correlation between tasks duration and subject's targets.

task	Spearman correlation coefficient	p-value
circA	0.7578	1.753E-13
circB	0.7029	4.719E-11
sigDiaA	0.6558	2.275E-09
sigDiaB	0.5746	4.526E-07
sigMea1	0.6758	4.796E-10

sigMea2	0.6415	6.491E-09
sigMea3	0.62	2.829E-08
sigMea4	0.5738	4.746E-07
sigSp1	0.6614	1.488E-09
sigSp2	0.6527	2.88E-09
sigSp3	0.6662	1.03E-09
sigSp4	0.612	4.763E-08
meanders	0.651	3.251E-09
spirals	0.6638	1.24E-09
all	0.7419	1.027E-12

2.2 Classification

To further confirm those results, I used sklearn implementation of Linear Discriminant Analysis (LDA). I feed the model only the duration of the exam (one feature).
In order to have comparable results with Afonso et al. I didn't use 10 CV but random 50-50 split with 15 runs. Pereira et al. evaluation method is unknown.

Accuracy (%) depending on the task (average over the 15 runs folds ± std).

task	LDA (my model)	Afonso et al.	Pereira et al. 2018
circA	80.61 ± (4.40)	76.17±6.92	68.04 ± 2.96
circB	83.43 ± (4.68)	76.69±5.38	73.41 ± 3.66
sigDiaA	75.96 ± (5.90)	68.69±7.26	73.59 ± 3.57
sigDiaB	77.17 ± (5.52)	66.30±7.38	76.32 ± 5.18
meanders	81.82 ± (3.13)	81.07±2.60	80.75 ± 2.08
spirals	82.02 ± (3.02)	81.03±2.40	78.26 ± 1.97
all	83.84 ± (3.26)	NA	95.74 ± 1.60

I averaged the lengths of the four meanders and the four spirals to provide the results of the lines meanders and spirals. Although I achieved similar results when using only one meander or one spiral to train the model. In the same way in "all" I averaged all the lengths for each subject. Majority voting provided similar results.
I displayed here the results of Afonso et al. when using first OPF then SVM because it's their best results (cf. Report #2). In the same way the results of Pereira et al. are the one where they used the "CNN-ImageNet" model on "images" of size 128x128 because it's their best results.
Notice how we outperform both of them on every single task although Pereira et al. achieved better results on all tasks (they use majority voting).

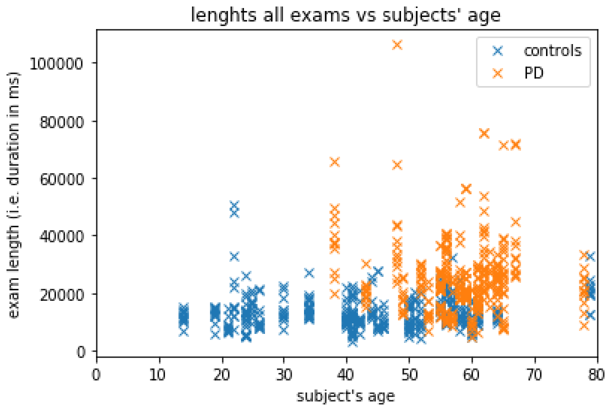
My goal here is not to outperform them but to prevent people from using this database which is obviously not challenging. Also it points out how one should be careful about age balance between patients and controls subjects.

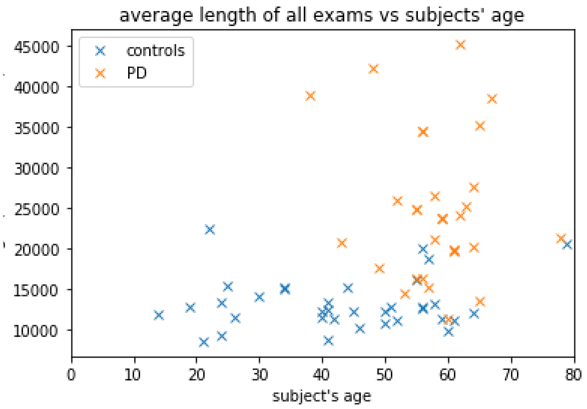
From that I think we can conclude that the transformation of the data from sensor to image of Pereira et al. is not good or that CNN are not suited to discriminate PD and one should focus on kinematic features. Also, it does not encourage the use of Discrete Wavelet Transform as did Afonso et al. (cf. Report #2).

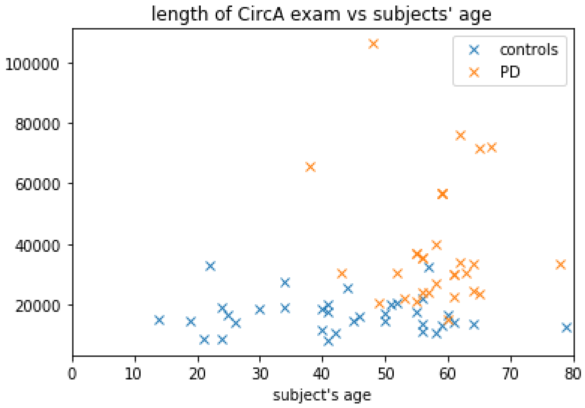
3 Interpretation

3.1 Age

The controls are in average 13.8 years younger than the PDs. In report #2 we saw that age had a lot of effect on dysgraphia and that most of the datasets aimed for a equally distributed age between PDs and control.
When plotting the exam length (i.e. duration) against the subject's age, we can see a nice cluster of PDs which partially explains why classification is so effective with only one feature. See below, the plot is more readable when plotting the average length per subject instead of every subjects' length. The phenomenon is visible on every single task, see below for the CircA exam.







I used Pearson's correlation coefficient which measures the linear correlation between two variables. It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

I used its SciPy implementation. to measure the correlation between the tasks duration and the age of the subject. As you can see below the durations of all tasks are positively correlated to the subject's age. The correlations of sigDia A and B are weaker than the other tasks. This explains why the classifier obtain worse results on these tasks (see 2.2 Classification). The correlation between task duration and subject's target was also weaker on sigDia B but not on A for some reason (see 2.1 Statistical analysis)...

Table. Correlation between tasks duration and subject's age.

task	pearson correlation coefficient	p-value	significance
circA	0.2781	0.0238	**
circB	0.2533	0.0401	**
sigDiaA	0.2099	0.0907	*
sigDiaB	0.2073	0.0948	*
sigMea1	0.3822	0.0015	***
sigMea2	0.3728	0.0021	***
sigMea3	0.3265	0.0075	***
sigMea4	0.3226	0.0082	***
sigSp1	0.3079	0.0119	**
sigSp2	0.3242	0.0079	***

sigSp3	0.3169	0.0095	***
sigSp4	0.3292	0.007	***
meanders	0.3619	0.0028	***
spirals	0.3346	0.006	***
all	0.3512	0.0038	***

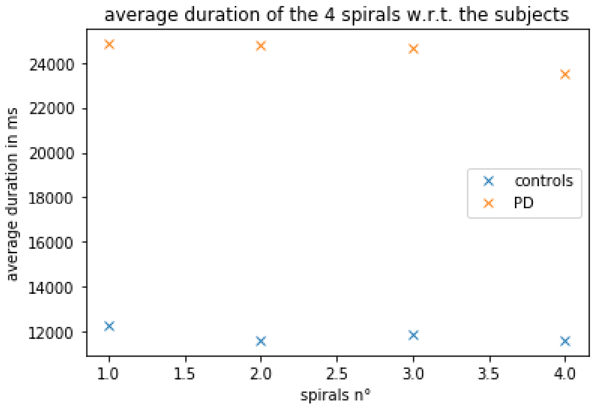
Significance : weak, strong and very strong for *, **, *** respectively

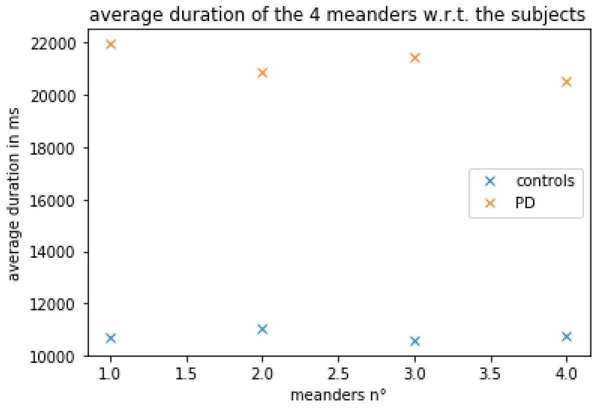
Moreover I used Spearman's rank correlation coefficient to measure the correlation between the subject's age and targets. It's **0.5359** with **p-value** 3.527e-06 (strongly significant).

Moreover, I tried to classify the subjects based only on their age, therefore using the same LDA model as before and achieved 75.76% accuracy ± 3.67% std over 15 runs of the 50-50 split.

3.2 Fatigue ?

One might ask if the duration of the exam is caused by fatigue and if PDs gets more tired than controls. Although proper statistical analysis might be required, when plotting the average duration of each 4 spirals and each 4 meanders in the order they were recorded, we cannot see any tendency (see below).





Conclusion - Todo List

I could have a look at the HandPD dataset to see if I obtain similar results.